

University of Dundee

SUPPA2

Trincado, Juan L.; Entizne, Juan C.; Hysenaj, Gerald; Singh, Babita; Skalic, Miha; Elliott, David J.

Published in:
Genome Biology

DOI:
[10.1186/s13059-018-1417-1](https://doi.org/10.1186/s13059-018-1417-1)

Publication date:
2018

Licence:
CC BY

Document Version
Publisher's PDF, also known as Version of record

[Link to publication in Discovery Research Portal](#)

Citation for published version (APA):

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*, 19, 1-11. [40]. <https://doi.org/10.1186/s13059-018-1417-1>

General rights

Copyright and moral rights for the publications made accessible in Discovery Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from Discovery Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain.
- You may freely distribute the URL identifying the publication in the public portal.

Take down policy


If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

METHOD

Open Access



SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions

Juan L. Trincado^{1†}, Juan C. Entizne^{2†}, Gerald Hysenaj³, Babita Singh¹, Miha Skalic¹, David J. Elliott³ and Eduardo Eyras^{1,4*} 

Abstract

Despite the many approaches to study differential splicing from RNA-seq, many challenges remain unsolved, including computing capacity and sequencing depth requirements. Here we present SUPPA2, a new method that addresses these challenges, and enables streamlined analysis across multiple conditions taking into account biological variability. Using experimental and simulated data, we show that SUPPA2 achieves higher accuracy compared to other methods, especially at low sequencing depth and short read length. We use SUPPA2 to identify novel Transformer2-regulated exons, novel microexons induced during differentiation of bipolar neurons, and novel intron retention events during erythroblast differentiation.

Keywords: Differential splicing, Alternative splicing, RNA-seq, Uncertainty, Biological variability, Differentiation

Background

Alternative splicing is related to a change in the relative abundance of transcript isoforms produced from the same gene [1]. Multiple approaches have been proposed to study differential splicing from RNA sequencing (RNA-seq) data [2, 3]. These methods generally involve the analysis of either transcript isoforms [4–7], clusters of splice junctions [8, 9], alternative splicing events [10, 11], or exonic regions [12]. Relative abundances of the splicing events or transcript isoforms are generally described in terms of a percentage or proportion spliced-in (PSI) and differential splicing is given in terms of the difference of these relative abundances, or Δ PSI, between conditions [13, 14]. PSI values estimated from RNA-seq data have shown a good agreement with independent experimental measurements, and the magnitude of Δ PSI represents a good indicator of biological relevance [10, 15]. However, despite the multiple improvements achieved by recent RNA-seq analysis methods, many challenges remain unresolved. These

include the limitations in processing time for current methods, the computational and storage capacity required, as well as the constraints in the number of sequencing reads needed to achieve high enough accuracy.

An additional challenge for RNA-seq analysis is the lack of robust methods to account for biological variability between replicates or to perform meaningful analyses of differential splicing across multiple conditions. Although many methods assess the estimation uncertainty of the splicing event or transcript isoforms [10–12], they generally do so on individual events rather than considering the genome-wide distribution. Additionally, most methods determine the significance of differential splicing by performing tests directly on read counts, leaving the selection of relevant Δ PSI values to an arbitrary cut-off. In other cases, fold changes instead of Δ PSI are given, which are even harder to interpret in terms of splicing changes.

We showed before that transcriptome quantification could be leveraged for the fast estimation of event PSI values with high accuracy compared with experimental and simulated datasets [16]. We now present here a new method for analyzing differential splicing, SUPPA2, which builds upon these principles to address the current challenges in the study of differential splicing, and taking into account biological variability. Compared

* Correspondence: eduardo.eyras@upf.edu

[†]Equal contributors

¹Pompeu Fabra University, E08003 Barcelona, Spain

⁴Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain

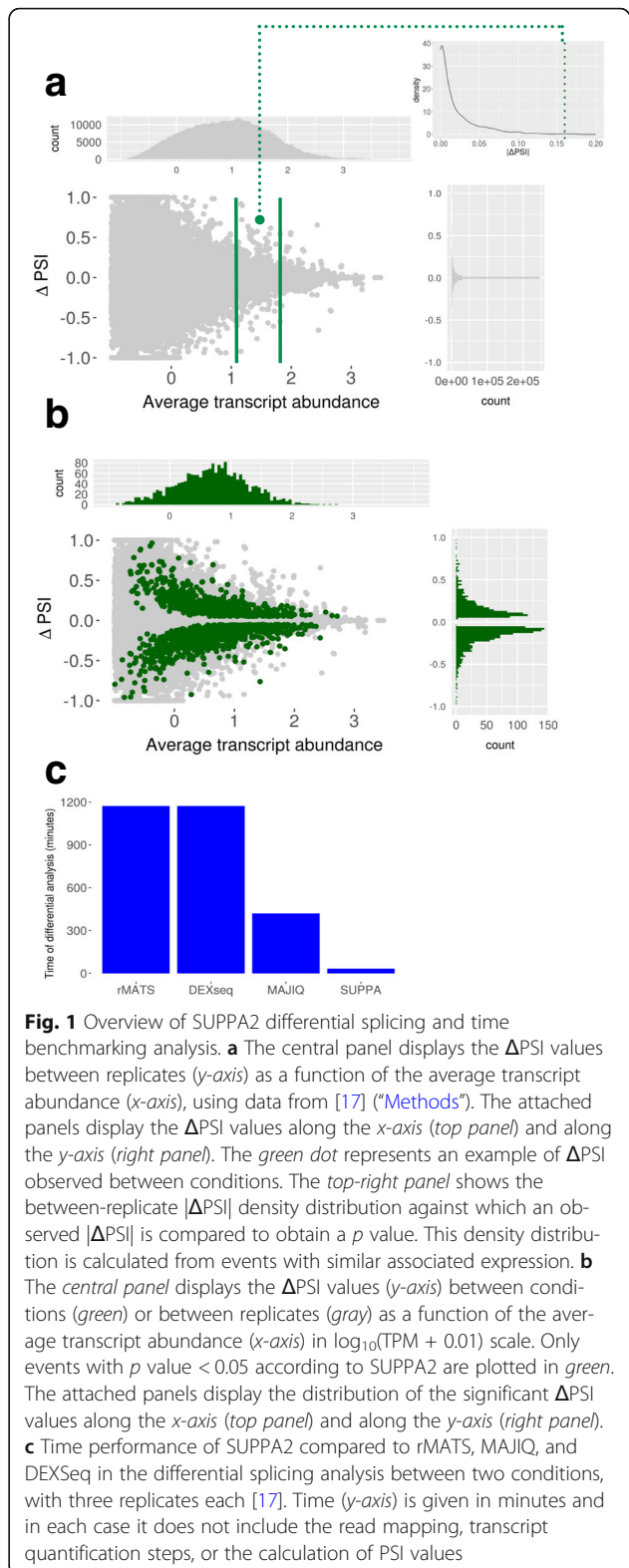
Full list of author information is available at the end of the article

with other existing approaches for differential splicing analysis using RNA-seq data, SUPPA2 provides several advantages. SUPPA2 can work with multiple replicates per condition and with multiple conditions. Additionally, SUPPA2 estimates the uncertainty of ΔPSI values as a function of the expression of transcripts involved in the event, taking into account all events genome-wide to test the significance of an observed ΔPSI , thereby directly estimating the biological relevance of the splicing change without relying on arbitrary ΔPSI cut-offs. Moreover, SUPPA2 incorporates the possibility to perform clustering of differentially spliced events across multiple conditions to identify groups of events with similar splicing patterns and common regulatory mechanisms. In conclusion, SUPPA2 enables cost-effective use of RNA-seq for the robust and streamlined analysis of differential splicing across multiple biological conditions. The software described here is available at <https://github.com/comprna/SUPPA>.

Results

SUPPA2 monitors uncertainty to determine differential splicing

We showed before that the inclusion levels of alternative splicing events can be readily calculated from transcript abundances estimated from RNA-seq data with good agreement with experimental measurements and with other methods based on local measurements of splicing [16]. SUPPA2 extends this principle to measure differential splicing between conditions by exploiting the variability between biological replicates to determine the uncertainty in the PSI values (see “Methods”). To illustrate our approach and to evaluate the dynamic range of SUPPA2 we used it to analyze RNA-seq data obtained after the double knockdown of TRA2A and TRA2B splicing regulators compared with controls [17] (Fig. 1a). The differences in PSI value for each event between biological replicates are higher at low expression, in agreement with the expected higher variability at low read count. This biological variability provides information on the uncertainty of the PSI estimates. The significance of an observed ΔPSI value between conditions will depend on where in the distribution of the uncertainty it falls. A large splicing change ($|\Delta\text{PSI}|$ value) may not be significant if it falls within a range of high uncertainty, whereas a small splicing change may be defined as robustly significant if it falls in the low uncertainty range. SUPPA2 estimates the significance considering the distribution between replicates for all events with similar transcript abundance; hence, it provides a lower bound for significant $|\Delta\text{PSI}|$ values that vary with the expression of the transcripts describing the event (Fig. 1b; see “Methods”). The description of the uncertainty in terms of transcript abundances, given in transcripts per million (TPM) units,



rather than read counts provides several advantages. These include speed, as there is no need to store or go back to read information, as well as interpretability and

application range, as transcript abundances are already normalized for transcript length and remain stable at different library sizes. More details on these advantages are provided below.

We compared SUPPA2 results with three other methods that calculate differential splicing using multiple replicates per condition: rMATS [11] and MAJIQ [9], which describe changes in terms of Δ PSI, and DEXSeq [12], which uses fold changes. Importantly, we found that SUPPA2 was much faster than the other methods, devoting 24 s to the PSI quantification and about 32 min and 47 s for differential splicing analysis on the same datasets (Fig. 1c). Since SUPPA2 performs the significance test directly on the Δ PSI values without needing to go back to the read data, it hence provides unmatched speed for differential splicing analysis. Comparing the results obtained with each method (Additional file 1: Figure S1), we observed that rMATS and DEXSeq detect many apparently significant events with small inclusion changes that are not distinguishable from the variability between biological replicates, whereas SUPPA2 and MAJIQ separate well these two distributions. As SUPPA2 exploits the between-replicate variability to test for significance, it avoids the use of an arbitrary global $|\Delta$ PSI| threshold to identify biologically relevant events and detects significant events across a wide range of gene expression values (Additional file 1: Figure S1). This feature of SUPPA2 should hence better rationalize $|\Delta$ PSI| threshold cut-offs.

SUPPA2 provides high accuracy at low sequencing depth and with short read lengths

To test the accuracy of SUPPA2 with different sequencing settings and compare it with other methods, we simulated 277 exon-cassette (SE) events and 318 alternative splice site (A5/A3) events with $|\Delta$ PSI| > 0.2 between two conditions with three replicates per condition (Additional file 1: Figure S2a). To perform a balanced comparison, we considered the same number of negative controls, consisting of different SE and A5/A3 events with arbitrary PSI values but with no simulated change between conditions (Additional file 2: Table S1; “Methods”). We simulated genome-wide RNA-seq reads using RSEM [18] at different sequencing depths (120, 60, 25, 10, and 5 million (M) 100-nucleotide (nt) paired-end reads per sample) and for different read lengths (100, 75, 50, and 25 nt at a fixed depth of 25 M paired-end reads). Despite the differences in the numbers and length of the reads (Additional file 2: Table S2), the genes containing the positive and negative events used for benchmarking showed similar distributions of expression values at all depths and read lengths (Additional file 1: Figure S2b). We then calculated differentially spliced events with SUPPA2, rMATS, MAJIQ, and DEXSeq and evaluated the detection rate and accuracy on the simulated events (Additional file 2: Table S3).

The detection rate was calculated as the proportion of simulated positive and negative cassette events that each method was able to measure from the RNA-seq data, i.e., the event was recovered regardless of whether it was detected as significant. The detection rate of SUPPA2 was superior than the other methods in all conditions, even at low depth and for shorter reads (Additional file 1: Figure S2c). We also measured the true positives, i.e., the positive events that were observed to change significantly and in the same direction by each method, and the false positives, i.e., the negative events predicted to change significantly. For SE events the true positive rates were comparable across different sequencing depths (Fig. 2a). On the other hand, for shorter read length SUPPA2 recovered a higher proportion of true positives compared to the other methods (Fig. 2b). For A5/A3 events we also observed a similar decay in true positives with sequencing depth for all methods (Fig. 2c) and a higher accuracy of SUPPA2 with shorter read lengths (Fig. 2d). The same accuracies were observed if we imposed in addition the cutoff $|\Delta$ PSI| > 0.2 for the predictions (Additional file 2: Table S3). The reduced proportion of true positives at low depth and shorter read length in other methods was probably due to them relying on having sufficient junction and/or exonic reads. Additionally, even though SUPPA2 recovered in general more negative events, i.e., events simulated to be not differentially spliced, the false positive rate remained comparable to the other methods, and below 5% for all conditions (Additional file 2: Table S3). To further evaluate the accuracies of the different methods, we computed receiver operating characteristic (ROC) and precision-recall (PR) curves (Additional file 2: Table S3). MAJIQ and SUPPA2 show similar areas under the ROC and PR curves, which drop at low depth and with short read lengths, whereas DEXSeq and rMATS show smaller areas across all values of depth and read length.

We also considered an unbalanced configuration where one replicate had 120 M reads and the other two replicates had 10 M reads. In this hybrid configuration, SUPPA2 recovered a high number of events and a high number of true positives for SE events. On the other hand, for A5/A3 events we observed a slight drop in accuracy (Additional file 2: Table S3), probably due to a high proportion of short variable regions in the alternative sites events (79 events (25%) of the A5/A3 events involved a region of under 9 nt), which may be more problematic for correct transcript quantification than using direct mapping to splice junctions. Importantly, although MAJIQ showed a high detection rate and accuracy in the unbalanced configuration, it had to be run with specialized parameters (“Methods”), whereas SUPPA2 was run in the same way for all cases. Additionally, SUPPA2 also showed high correlation values between

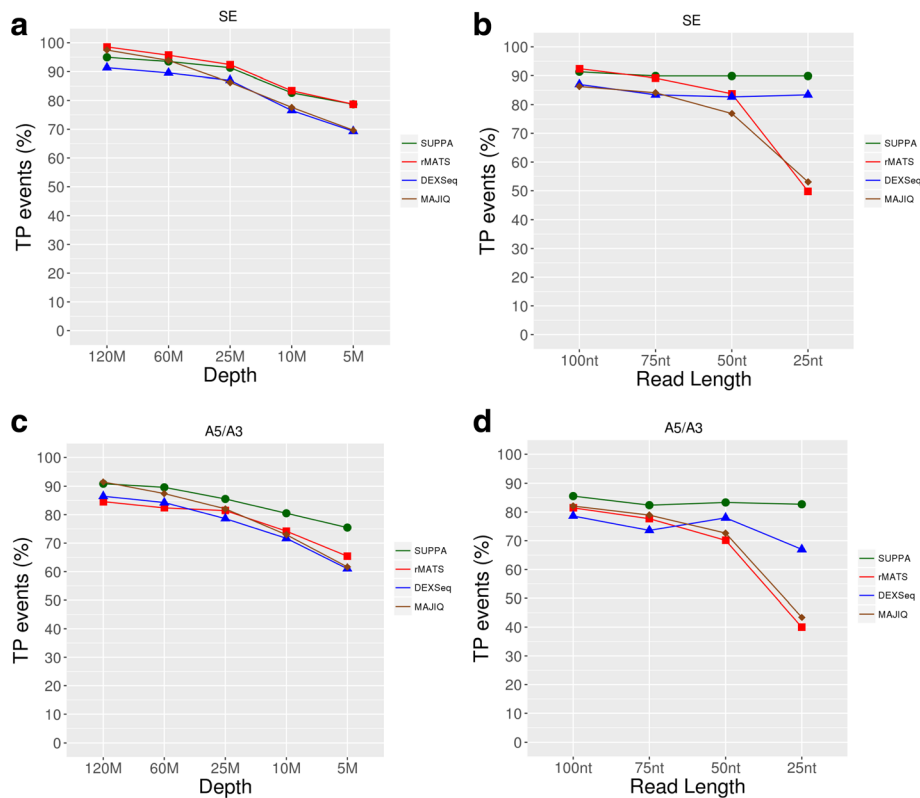


Fig. 2 Accuracy analysis with simulated data. **a** Proportion of events measured by each method (*y-axis*) from the 277 positive simulated cassette events at different sequencing depths (*x-axis*), from 120 million (120M) down to five million (5M) paired-end reads, using 100-nt paired-end reads. **b** As in **a** but for different read lengths (*x-axis*) at fixed depth (25 M). **c** True positive (TP) rate (in terms of percentage) for each method (*y-axis*) at different sequencing depths (*x-axis*) for 100-nt paired-end reads. TPs were calculated as the number of statistically significant events according to each method: corrected p value < 0.05 for SUPPA2, rMATS, and DEXSeq; and posterior($|\Delta\text{PSI}| > 0.1$) > 0.95 for MAJIQ. **d** As in **c** but for different read lengths (*x-axis*) at fixed depth (25 M)

the predicted and simulated ΔPSI values (Additional file 2: Table S3), and similar to those obtained with rMATS and MAJIQ. In light of these results, we can conclude that SUPPA2 performs comparably to other methods under a wide spectrum of sequencing conditions and, in particular, it outperforms other methods at low sequencing depth and short read length.

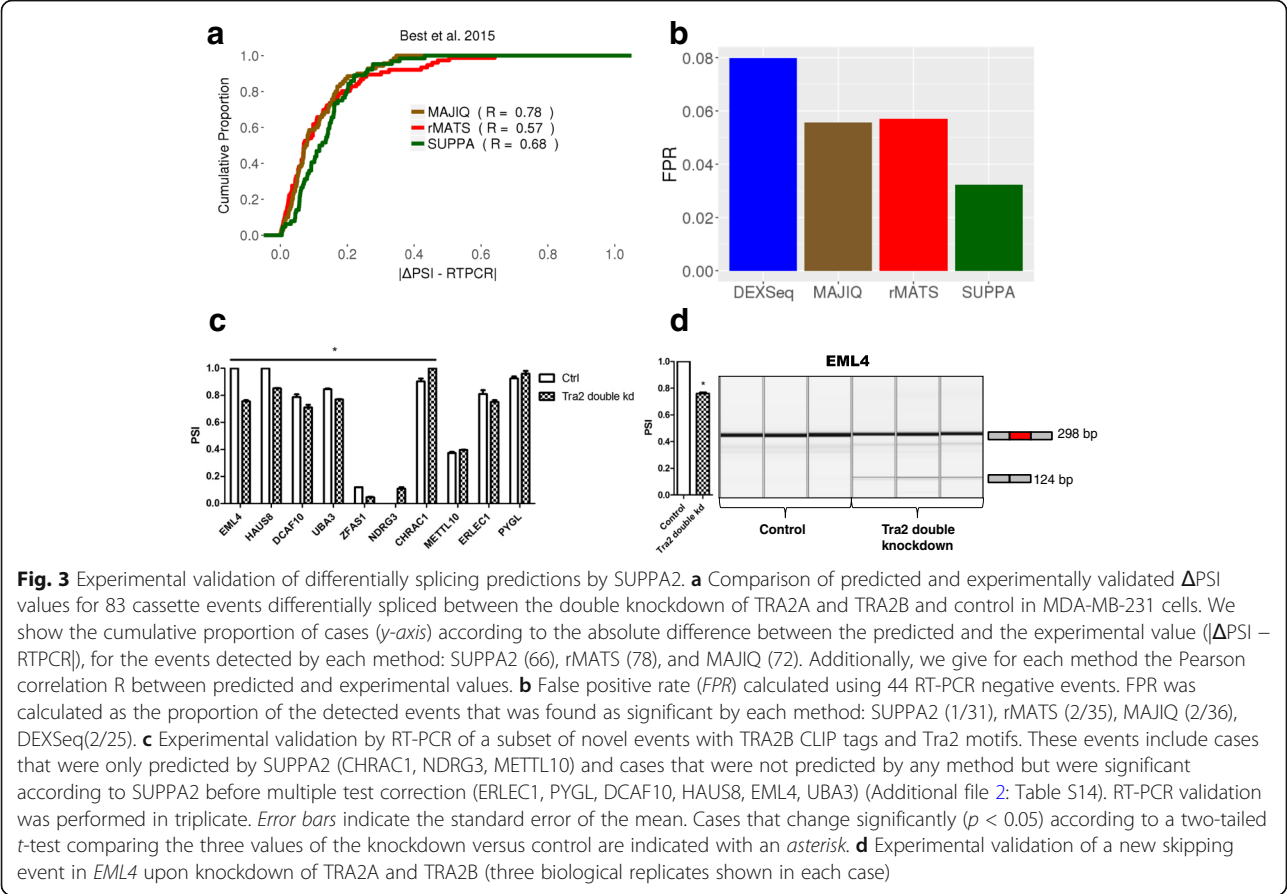
SUPPA2 provides accurate splicing change quantification compared with experimental results

To further evaluate the accuracy of SUPPA2 in recovering ΔPSI values we used 83 events that had been validated experimentally by RT-PCR upon TRA2A and TRA2B knockdown compared to control cells (Additional file 2: Table S4; “Methods”) [17]. For each method, we compared the ΔPSI estimated from RNA-seq with the ΔPSI from RT-PCR. SUPPA2 agreement to the RT-PCR ΔPSI values was similar to rMATS and MAJIQ (Fig. 3a; Additional file 2: Table S5). Using two other independent RT-PCR datasets published previously [9], SUPPA2 also showed similar accuracy compared to rMATS and MAJIQ (Additional file 1: Figure S3a, b; Additional file 2: Tables S6–S9). Finally, using

44 RT-PCR negative cassette events that did not show any significant change upon the double knockdown of TRA2A and TRA2B, SUPPA2 had a lower false positive rate compared to the other methods (Fig. 3b; Additional file 2: Tables S10 and S11).

SUPPA2 identifies experimentally reproducible splicing changes not detected by other methods

The results described above suggest a general agreement between the different methods in the detection of significant differentially spliced events. To assess this question, we performed a direct comparison of the results obtained from the four methods, SUPPA2, rMATS, MAJIQ, and DEXSeq, using the same RNA-seq data for the knockdown of TRA2A and TRA2B compared with controls [17]. Since exon-cassette (SE; 48.71%) and alternative splice site (A5/A3; 37.71%) events are the most frequent events in humans compared to mutual exclusion (6.22%) or intron-retention (7.36%), we decided to match SE and A5/A3 events across all four methods. We were able to identify 7116 SE events and 2924 A5/A3 events unambiguously detected by all four methods,



i.e., they were measured and tested for significance by all methods (Additional file 1: Figure S4a; Additional file 2: Table S12; “Methods”).

For the 7116 SE events, each method found between 133 and 274 events to be significant, with 370 events predicted as significant by any one method, but only 22 events predicted by all four methods (Additional file 1: Figure S4a). Similarly, 352 A5/A3 events were predicted to be significant by at least one method, and only two predicted by all four methods (Additional file 1: Figure S4a). Events detected by more methods tended to have higher Δ PSI values (Additional file 1: Figure S4b) and covered a smaller range of gene expression values (Additional file 1: Figure S4c). Despite the low detection overlap, the significant events predicted by each method independently showed enrichment of TRA2B CLIP tags and of Tra2 binding motifs (Additional file 2: Table S13; Additional file 3: Supplementary methods); hence, each set independently had the expected properties related to the knockdown experiment. It is possible that each method describes a different subset of changes and generally misses others. To seek further support for this point, we selected for experimental validation 15 SE events and seven A3 events that had CLIP tags and Tra2 motifs nearby the regulated exon. The seven A3 events and six of the 15 SE events were predicted only

by SUPPA2, whereas the remaining nine were not predicted by any of the four methods, but were significant according to SUPPA2 before multiple test correction (Additional file 2: Table S14). From these 15 SE events, five only showed one PCR band and could not be evaluated. However, for the rest, seven changed significantly according to the RT-PCR (two-tailed t -test p value < 0.05), with six of them changing in the same direction predicted by SUPPA2. Overall, nine events changed in the same direction as predicted (Fig. 3c; Additional file 2: Table S14). In particular, we validated a new event in *EML4* (Fig. 3d), a gene involved in cancer through a fusion with *ALK* that is not present in MDA-MB-231 cells [18]. In addition, we could measure six of the seven A3 events; all were measured to change in the same direction as predicted by SUPPA2 and four were significant (two-tailed t -test p value < 0.05 ; Additional file 2: Table S14). This analysis shows the value of using a suite of methods based on different algorithms, like SUPPA2, to reveal novel experimentally reproducible events that are missed by other methods.

SUPPA2 finds biologically relevant event clusters across multiple conditions

SUPPA2 is also able to analyze multiple conditions by computing the pairwise differential splicing between conditions,

and can detect groups of events with similar splicing patterns across conditions using density-based clustering (“Methods”). To evaluate the ability of SUPPA2 to cluster events, we analyzed a 4-day time-course of differentiation of human induced pluripotent stem cells (iPSCs) into bipolar neurons [19], which had not been analyzed yet for alternative splicing. SUPPA2 identified 2780 regulated cassette events (p value < 0.05), out of which 207 (8.4%) were microexons (length < 28 nt), which represent an enrichment (Fisher’s exact test p value $< 2.2\text{e-}16$, odds ratio = 3.94) compared to a set of 20,452 non-regulated cassette events (p value > 0.1), with the majority of these microexons (69%) significantly more included in differentiated cells ($\Delta\text{PSI} > 0$ and p value < 0.05 between the first and fourth day).

We evaluated the performance of the two density-based cluster methods implemented in SUPPA2, DBSCAN [20], and OPTICS [21], using different input parameters. In spite of OPTICS requiring more computing time than DBSCAN (43 vs 5 s), it produced slightly better clustering results (Additional file 1: Figure S5a–d; Additional file 2: Table S15). For a maximum reachability distance of 0.11, i.e., maximum distance of an event to a cluster to be

considered part of the cluster, we obtained three well-differentiated clusters (silhouette score = 0.572; Fig. 4a–c; Additional file 2: Table S16). Cluster 0 increased inclusion at late steps of differentiation and showed an enrichment in microexons (32 out of 115 events) with respect to unclustered regulated cassette events (Fisher’s exact test p value = 0.0148, odds ratio = 5.3521). In contrast, clusters 1 and 2 decreased inclusion with differentiation, and contained two (out of 20 events) and no microexons, respectively. These results are in agreement with the previously observed enrichment of microexon inclusion in differentiated neurons [22, 23].

To further validate the findings with SUPPA2, we performed a motif enrichment analysis in regulated events compared to non-regulated events. Notably, compared to the non-regulated events, the 2780 regulated cassette events showed enrichment in binding motifs for the RNA binding protein (RBP) SFPQ (z-score > 4), which has been described before as a necessary factor for neuronal development [24]. Additionally, the differentially spliced events in clusters were enriched in, among others, CELF, RBFOX, ESRP, MBNL, and SRRM4 motifs

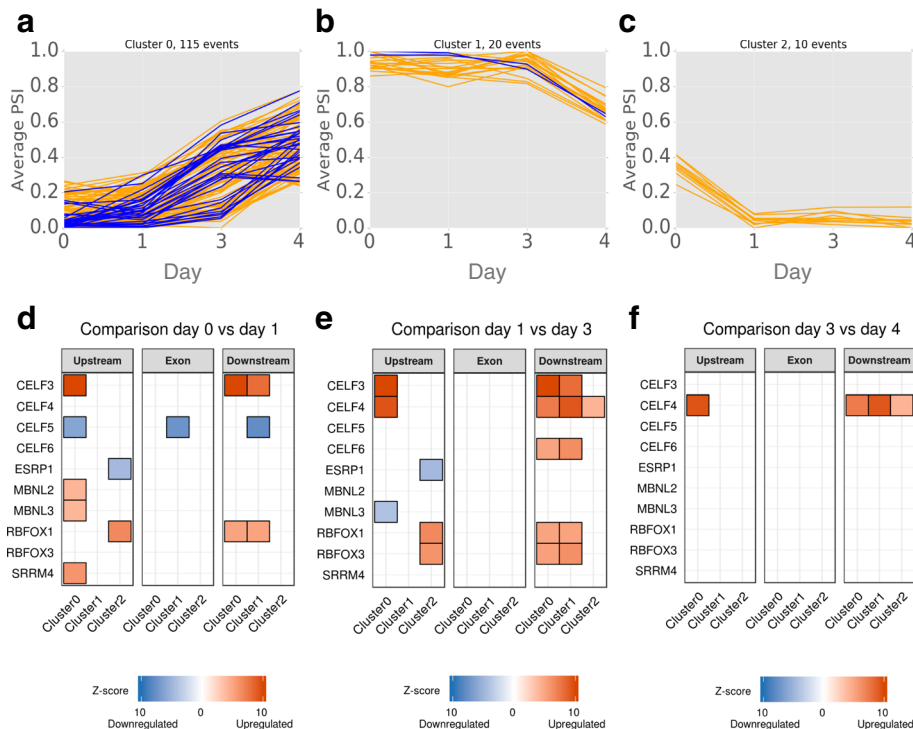


Fig. 4 Prediction and clustering of differentially spliced events across bipolar neuron differentiation. Density-based clustering performed on the 2780 regulated cassette events that change splicing significantly in at least one comparison between adjacent steps across four differentiation stages (days after differentiation 0, 1, 3, 4). **a–c** The average PSI (y-axis) per stage (x-axis) of the events in the three clusters obtained. Microexons (< 28 nt) are plotted in blue over the rest of the events in orange. **d–f** Motif enrichment associated with each of the three clusters in **a–c** in the regions upstream (200 nt), exonic, and downstream (200 nt). Only enriched motifs associated with splicing factors that are differentially expressed are shown in each comparison between differentiation stages (days after differentiation 0, 1, 3, 4). In red we indicate the splicing factors that are upregulated and in blue those that are downregulated at each stage. The color intensity indicates the z-score of the motif enrichment. Motifs are shown in each cluster and region where they are found enriched

(Fig. 4d–f), in concordance with the described role of *CELF*, *RBFOX*, and *SRRM4* genes in neuronal differentiation [23, 25–27]. Consistent with these findings, *SRRM4* and members of the *CELF* and *RBFOX* families showed upregulation at the initial steps of iPSC differentiation into neurons (Additional file 1: Figure S5; Additional file 2: Table S17). On the other hand, *CELF5* and *ESRP1* were downregulated during differentiation. The *MBNL3* gene showed initial upregulation at stage 1, followed by downregulation at later stages (Additional file 1: Figure S5; Additional file 2: Table S17). Notably, we found that only the cluster enriched in microexon splicing inclusion showed an enrichment of *SRRM4* motifs upstream of the regulated exons, in agreement with the previous description of *SRRM4* binding upstream of microexons to regulate their inclusion during neuronal differentiation [26], and further supports the specificity of *SRRM4* to regulate microexons. Our results also suggest possible novel regulators of neuronal differentiation, such as the *MBNL* proteins in the regulation of events increasing exon inclusion and *ESRP* in events that decrease exon inclusion (Fig. 4d–f).

We also used SUPPA2 to analyze differential splicing across five stages of erythroblast differentiation [28]. In this case we considered all event types for clustering. For the optimal value of maximum reachability distance ($S = 0.1$), we obtained two homogeneous and well-differentiated clusters (silhouette score = 0.91), one for events with low PSI that increased at the last differentiation stage with 149 events, and a second cluster with 86 events that showed the opposite behavior (Additional file 1: Figure S6). In agreement with previous results [29], we observed an enrichment of intron retention events in the cluster of events that increased inclusion at the late differentiation stage, as compared with the other cluster, which does not include any retained intron (Fisher's exact test p value = 0.04958). We conclude that SUPPA2 provides a powerful approach to analyze splicing across multiple conditions, validated not only by intrinsic measures of clustering consistency, but also by recovering known biological results and new features.

Discussion

Our extensive evaluations here indicate that SUPPA2 provides a broadly applicable solution to current challenges in the analysis of differential splicing from RNA sequencing data across multiple conditions, and has features that will make it attractive to many potential users. SUPPA2 is faster than other methods and maintains a high accuracy, especially at low sequencing depth and for short read length. Despite using less reads or shorter reads, SUPPA2 could detect the majority of the simulated events and maintained a high proportion of true positives and low proportion of false positives.

SUPPA2 thus offers an unprecedented opportunity to study splicing in projects with limited budgets, or to re-use for splicing studies available sequencing datasets with lower depth than usually required by other methods. Additionally, the low computing and storage requirements of SUPPA2 makes it possible to perform fast differential splicing processing and clustering analysis on a laptop. Thus, coupled with fast methods for transcript quantification [30–32], SUPPA2 facilitates the study of alternative splicing across multiple conditions without the need for large computational resources. The simplicity and modular architecture of SUPPA2 also makes it a very convenient tool in multiple contexts, as PSI values from other methods and for other event types, like complex events, or data types, like transcripts, can be used in SUPPA2 for differential splicing analysis or for clustering across conditions.

According to our simulated benchmarking analysis, as well as others published before, it may seem that bioinformatics methods used to analyze RNA-seq data tend to coincide on a large number of events. However, using real experimental data we actually observed low agreement in targets between methods. These discrepancies in target selection can be explained by various factors, including the different ways in which a splicing change is represented by each method (e.g., an event, an exon, or a graph), how changes in splicing patterns are tested by each method, and how biological and experimental variability affects these tests. Intriguingly, the results from each method do make sense biologically, in that differentially spliced events were enriched in motifs and mapped protein–RNA interaction sites related to the depleted splicing factor. This makes it unlikely that any one method provides a clear advantage in terms of the results, and instead suggests that at least two or three methods should be used to identify all the possible significant splicing variants between different conditions. In particular, we chose for comparison three other methods with very different representations of the splicing and statistical approach. The results we obtained recommend use of two or more such tools to comprehensively monitor splicing complexity by picking out different sets of events that would not otherwise be discovered, rather than identifying largely overlapping groups of events. Supporting this point we could validate experimentally events not predicted by any other methods but predicted by SUPPA2. We further observed that although most methods had the power to identify small significant Δ PSI values, different methods tended to agree on events with large splicing changes. Importantly, a fraction of these significant events with small Δ PSI are indistinguishable from the variability observed between replicates and hence are not likely to be biologically relevant. SUPPA2 also performs a statistical test

that can separate significant splicing changes from the biological variability, providing thus an advantage to identify biologically relevant changes across a wide range of expression values. By exploiting the biological variability, without having to go back to the read data, SUPPA2 provides a fast and accurate way to detect differential splicing without the need for arbitrary global Δ PSI thresholds.

Although SUPPA2 relies on genome annotation to define events, poorly annotated genomes can be improved and extended before analysis by SUPPA2. In fact, recent analyses have shown that improved annotations lead to significantly better PSI estimates from RNA-seq when benchmarked to high-resolution RT-PCR measurements [33–35]. Current technological trends predict an increase in the number of efforts to improve the transcriptome annotation in multiple species and conditions [36]. In this direction, SUPPA2 could play a key role for the systematic and rapid genome-wide analysis of splicing following annotation and sample updates.

Conclusions

The speed, modularity, and accuracy of SUPPA2 enable cost-effective use of RNA sequencing for the robust and streamlined analysis of differential splicing across multiple biological conditions.

Methods

Differential splicing

SUPPA2 uses transcript quantification to compute inclusion values (PSI) of alternative splicing events across multiple samples. Given the calculated PSI values per sample, SUPPA2 considers two distributions: one for the Δ PSI values between biological replicates and one for the Δ PSI values between conditions. For the first distribution, for each event SUPPA2 calculates the Δ PSI value between each pair of biological replicates together with the average abundance of the transcripts describing the event across the same replicates:

$$E_{rep} = \frac{1}{|R_c|} \sum_{r \in R_c} \log_{10} \left(\sum_a TPM_{a,r} \right)$$

where $r = 1, \dots, |R_c|$ runs over the replicates in each condition $c = 1, 2$, and a indicates the two or more transcripts describing the event, and $TPM_{a,r}$ indicates the abundance of transcript a in replicate r in transcripts per million (TPM) units. For the distribution between conditions, the Δ PSI values are calculated as the difference of the means in the two conditions, together with the average abundance of transcripts describing the event across both conditions for each event:

$$E_{cond} = \frac{1}{2} \sum_{c=1,2} \frac{1}{|R_c|} \sum_{r \in R_c} \log_{10} \left(\sum_a TPM_{a,r,c} \right)$$

where $TPM_{a,r,c}$ indicates the abundance of transcript a in replicate r in condition c in TPM units. Given the observed Δ PSI and E_{cond} values for an event between conditions, its significance is calculated from the comparison with the Δ PSI distribution between replicates for events with E_{rep} values in the neighborhood of the observed E_{cond} . This neighborhood is defined by first selecting the closest value E_{rep}^* from all points i from the between-replicate distribution:

$$E_{rep}^* = \min_i \{ |E_{i,rep} - E_{cond}| \}$$

using binary search and selecting a fixed number of events (1000 by default) around the E_{rep}^* value in the interval or ordered values. The selected events define an empirical cumulative density function (ECDF) over $|\Delta$ PSI| from which a p value is calculated:

$$p = (1 - ECDF(|\Delta$$
PSI|))/2

Here we implicitly assume that the background distribution is symmetric. SUPPA2 includes an option to correct for multiple testing using the Benjamini-Hochberg method across all events from the same gene, as they cannot be considered to be entirely independent of each other, for which the false discovery rate (FDR) cut-off can be given as input.

Clustering

SUPPA2 currently implements two density-based clustering methods: DBSCAN [20] and OPTICS [21]. Density-based clustering has the advantage that one does not need to specify the expected number of clusters, and the choice between the two methods depends mainly on the computational resources and the amount of data. Both methods use the vectors of mean PSI values per event and require as input the minimum number of events in a cluster (N), which could be interpreted as the minimum expected size of the regulatory modules. OPTICS also requires the maximum reachability distance (S), which represents the maximum distance in PSI space of an event to a cluster. On the other hand, DBSCAN requires as input the maximum distance to consider two events as cluster partners (D), which OPTICS calculates through an optimization procedure allowing any value below S . DBSCAN allows simple and fast data partitioning but has the drawback of being sensitive to the input parameters. On the other hand, OPTICS, which can be seen as a generalization of DBSCAN, explores the possible maximum values for D beyond which clustering quality drops. OPTICS can thus

potentially produce better clustering results since it is not limited to a fixed radius of clustering, but it is penalized by a greater computational cost. Clustering is performed only with events that change significantly in at least one pair of adjacent conditions. Three different distance metrics can be currently used: Euclidean, Manhattan, and Cosine. Cluster qualities are reported using the silhouette score [37], which indicates how well the events are assigned to clusters, and the root mean square standard deviation (RMSSTD), which measures the homogeneity of each cluster. Additionally, the number and percentage of events in clusters are also reported. Motif enrichment analysis was performed as before [38] using MOSEA, available at <https://github.com/comprna/MOSEA>. Further details on the motif enrichment and analysis of differential expression are provided in Additional file 3: Supplementary material.

Simulated datasets

For the simulation we used the quantification of RefSeq transcripts for the three control samples from [17] (GSE59335) with Salmon [31] as theoretical abundances, and considered genes with only two isoforms containing a skipping exon (SE) or alternative splice site (A5/A3) event and only one associated event. For the benchmarking analysis, we selected a set of positive and a set of negative events for each event type with the same number of randomly chosen events, 277 for SE events and 318 for A5/A3 events. For the positive set we simulated differential splicing by exchanging the theoretical abundance of their associated transcript values. We selected to be positive events only those having an absolute difference of relative abundance greater than 0.2, so that the simulated change was sufficiently large:

$$\frac{|TPM_1 - TPM_2|}{TPM_1 + TPM_2} > 0.2$$

where TPM1 and TPM2 are the abundances for the two transcripts in the gene, given in TPM units. For the negative set, we took an equal number of events without exchanging their TPM values. These negative events had a gene expression distribution and a distribution of transcript relative abundance similar to the positive events, and an expected variability between conditions similar to the variability between biological replicates. We used RSEM [39] to simulate sequencing reads for the two conditions, three replicates each, at various depths (120, 60, 25, 10 and 5 M 100-nt paired-end reads per sample) and at various read lengths (100, 75, 50, and 25 nt, at a depth of 25 M paired-end reads) (Additional file 2: Tables S1–S3). Further details of the simulations are given in the Additional file 3: Supplementary material. Datasets and

commands to reproduce these simulations are available at https://github.com/comprna/SUPPA_supplementary_data.

Experimental datasets

We analyzed RNA-seq data for the double knockdown of TRA2A and TRA2B in MDA-MB-231 cells and controls with three replicates per condition [17] (GSE59335). For benchmarking, we used 83 RT-PCR validated events for comparison (Additional file 2: Tables S4 and S5) and 44 RT-PCR negative events (Additional file 2: Tables S12 and S13). We also analyzed data from cerebellum and liver mouse tissues covering eight different time points from two full circadian cycles [40] (GSE54651) and performed a comparison with 50 events validated by RT-PCR [9] comparing samples CT28, CT40, and CT52 in cerebellum with the same circadian time points in liver (Additional file 2: Tables S8 and S9). We also analyzed RNA-seq data for stimulated and unstimulated Jurkat T cells and compared them with RT-PCR validated events (no tested replicates) [9, 41] (SRP059357; Additional file 2: Tables S10 and S11). From these 54 RT-PCR validated events, we only used the 30 events that had experimental value $|\Delta PSI| > 0.05$. For the study of multiple conditions, we used RNA-seq samples from a 4-day time-course for the differentiation of human iPSCs into bipolar neurons [19] (GSE60548). Original data were for days 0, 1, 3, and 4 after initiation of differentiation. Additionally, we analyzed RNA-seq from five steps of differentiating human erythroblasts [29] (GSE53635), with three replicates per condition. RNA-seq reads from all experiments were used to quantify human and mouse transcripts from Ensembl (version 75, without pseudogenes) with Salmon [31]. Reads were mapped to the human (hg19) or mouse (mm10) genomes using TopHat [42]. All methods other than SUPPA2 were used with these mappings. Cassette events from SUPPA2 and rMATS were matched to the RT-PCR validated events in each dataset, considering only those cases where the middle exon matched exactly the validated exons and confirming the flanking exons with the RT-PCR primers when available. Ambiguous matches were discarded from the comparison. For MAJIQ we selected the inclusion junction compatible with the validated event that had the largest posterior probability for $|\Delta PSI| > 0.1$. For DEXSeq we considered only exonic regions that matched exactly with the regulated exon of the experimentally validated cassette event. To select a set of cassette events common to all four methods, we selected the events measured by both SUPPA2 and rMATS such that the middle exon matched exactly a DEXSeq exonic region and did not appear in more than one event from SUPPA2 or rMATS. From this set, we selected those for which any of the two inclusion junctions was present in MAJIQ, and selected the junction with the largest posterior probability for $|\Delta PSI| > 0.1$. Further details are provided in Additional file 3: Supplementary material.

Time performance

Running time was measured using the Unix time command *time*. For SUPPA2 running time was measured independently of the transcript quantification step. Similarly, for all other methods the running time did not include the read-mapping step. Time was measured independently for PSI calculation and for differential splicing analysis. All methods were run on a Unix machine with 12 Gb of RAM and eight Intel Xeon 2-GHz CPU cores.

Experimental validation

Details on the experimental validation are given in Additional file 3: Supplementary material.

Software and datasets

SUPPA2 is available at <https://github.com/comprna/SUPPA>.

Commands and datasets used in this work are available at https://github.com/comprna/SUPPA_supplementary_data.

Software for the motif enrichment analysis is available at <https://github.com/comprna/MOSEA>.

Additional files

Additional file 1: Figures S1–S6. (PDF 3939 kb)

Additional file 2: Tables S1–S17. (XLSX 3050 kb)

Additional file 3: Supplementary methods. (PDF 315 kb)

Abbreviations

CLIP: Cross-linking immunoprecipitation; iPSC: Induced pluripotent stem cell; PSI: Proportion spliced in; RNA-seq: RNA sequencing; RT-PCR: Reverse transcriptase polymerase chain reaction; TPM: transcripts per million; TRA2A/B: Transformer-2 protein homolog alpha/beta

Acknowledgements

The authors thank C. Calixto, J. Brown, R. Zhang, M. Irimia, and N. Barbosa-Morais for useful discussions and J. Vaquero-Garcia and Y. Barash for comments on an earlier version of the manuscript.

Funding

This work was supported by the MINECO and FEDER with grants BIO2014–52566-R and BIO2017–85364-R, by AGAUR with grants SGR2014–1121 and SGR2017–1020, by BBSRC (BB/P006612/1), and by Breast Cancer Now (2014NovPR355). GH is a BBSRC-funded PhD student.

Availability of data and materials

Datasets used in this manuscript are accessible at the Gene Expression Omnibus (GEO) under accessions GSE59335 [43], GSE54651 [44], GSE60548 [45], and GSE53635 [46]. SUPPA (version 2.2 [47]) is open source under the MIT license (OSI-compliant), is implemented in python3.4, and is freely available at <https://github.com/comprna/SUPPA>.

Authors' contributions

JCE, MS, JLT, and EE designed and implemented the method and algorithms, JCE, JLT, and EE devised the analyses, and JCE, JLT, and MS carried out the benchmarking analyses. GH and DJE produced the datasets related to the double knockdown of TRA2A and TRA2B and performed the validation experiments. BS developed the software for motif enrichment analysis. JLT, JCE, and EE wrote the manuscript with essential input from GH and DJE. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Ethics approval was not applicable for this study.

Competing interests

The authors declare that they have no competing interests.

Author details

¹Pompeu Fabra University, E08003 Barcelona, Spain. ²University of Dundee, Invergowrie, Dundee DD2 5DA, UK. ³Institute of Genetic Medicine, Newcastle University, Central Parkway, Newcastle NE1 3BZ, UK. ⁴Catalan Institution for Research and Advanced Studies, E08010 Barcelona, Spain.

Received: 30 November 2017 Accepted: 2 March 2018

Published online: 23 March 2018

References

- Lee Y, Rio DC. Mechanisms and regulation of alternative pre-mRNA splicing. *Annu Rev Biochem*. 2015;84:291–323. <https://www.ncbi.nlm.nih.gov/pubmed/25784052>.
- Alamancos GP, Agirre E, Eyra E. Methods to study splicing from high-throughput RNA sequencing data. *Methods Mol Biol*. 2014;1126:357–97. <https://www.ncbi.nlm.nih.gov/pubmed/24549677>.
- Lahat A, Grellscheid SN. "Differential mRNA Alternative Splicing." In *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*. Cham: Springer; 2016. p. 105–119. https://link.springer.com/chapter/10.1007/978-3-319-31350-4_5.
- Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013;31:46–53. <http://dx.doi.org/10.1038/nbt.2450>
- Sebestyén E, Zawisza M, Eyra E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. *Nucleic Acids Res*. 2015;43:1345–56. <http://www.ncbi.nlm.nih.gov/pubmed/25578962>
- Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research*. 2016;5:1356. <http://www.ncbi.nlm.nih.gov/pubmed/28105305>
- Froussios K, Mourão K, Schurch NJ, Barton GJ. Identifying differential isoform abundance with RATs: a universal tool and a warning. *bioRxiv*. 2017. p.132761.
- Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Res*. 2013;41(2):e39. <https://www.ncbi.nlm.nih.gov/pubmed/23155066>.
- Vaquero-Garcia J, Barrera A, Gazzara MR, González-Vallinas J, Lahens NF, Hogenesch JB, et al. A new view of transcriptome complexity and regulation through the lens of local splicing variations. *elife*. 2016;5:e11752. <http://www.ncbi.nlm.nih.gov/pubmed/26829591>
- Katz Y, Wang ET, Airolidi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods*. 2010;7:1009–15. <http://www.ncbi.nlm.nih.gov/pubmed/21057496>
- Shen S, Park JW, Lu Z, Lin L, Henry MD, Wu YN, et al. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-seq data. *Proc Natl Acad Sci U S A*. 2014;111:E5593–601. <http://www.ncbi.nlm.nih.gov/pubmed/25480548>
- Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-seq data. *Genome Res*. 2012;22:2008–17. <http://www.ncbi.nlm.nih.gov/pubmed/22722343>
- Wang ET, Sandberg R, Luo S, Khrebukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008;456:470–6. <http://www.ncbi.nlm.nih.gov/pubmed/18978772>
- Venables JP, Klinck R, Bramard A, Inkel L, Dufresne-Martin G, Koh C, et al. Identification of alternative splicing markers for breast cancer. *Cancer Res*. 2008;68:9525–31. <http://www.ncbi.nlm.nih.gov/pubmed/19010929>
- Venables JP, Brosseau J-P, Gadea G, Klinck R, Prinos P, Beaulieu J-F, et al. RBFOX2 is an important regulator of mesenchymal tissue-specific splicing in both normal and cancer tissues. *Mol Cell Biol*. 2013;33:396–405. <https://www.ncbi.nlm.nih.gov/pubmed/23149937>
- Alamancos GP, Pagés A, Trincado JL, Bellora N, Eyra E. Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA*. 2015;21:1521–31. <https://www.ncbi.nlm.nih.gov/pubmed/26179515>
- Best A, James K, Dalgliesh C, Hong E, Khairilahi-Kouhestani M, Curk T, et al. Human Tra2 proteins jointly control a CHEK1 splicing switch among alternative and constitutive target exons. *Nat Commun*. 2014;5:4760. <http://www.ncbi.nlm.nih.gov/pubmed/25208576>

18. Lin E, Li L, Guan Y, Soriano R, Rivers CS, Mohan S, et al. Exon array profiling detects EML4-ALK fusion in breast, colorectal, and non-small cell lung cancers. *Mol Cancer Res*. 2009;7:1466–76. <http://www.ncbi.nlm.nih.gov/pubmed/19737969>
19. Busskamp V, Lewis NE, Guye P, Ng AHM, Shipman SL, Byrne SM, et al. Rapid neurogenesis through transcriptional activation in human stem cells. *Mol Syst Biol*. 2014;10:760. <http://www.ncbi.nlm.nih.gov/pubmed/25403753>
20. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*; 1996. p. 226–31. <https://www.aaii.org/Papers/KDD/1996/KDD96-037.pdf>
21. Ankerst M, Breunig MM, Kriegel H, Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. *ACM SIGMOD Rec*. 1999;28:49–60.
22. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell*. 2014;159:1511–23. <http://www.ncbi.nlm.nih.gov/pubmed/25525873>
23. Li Yi, Sanchez-Pulido L, Haerty W, Ponting CP. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts. *Genome Res*. 2015;25:1–13. <http://www.ncbi.nlm.nih.gov/pubmed/25524026>
24. Lowery LA, Rubin J, Sive H. Whitesnake/sfpq is required for cell survival and neuronal development in the zebrafish. *Dev Dyn*. 2007;236:1347–57. <http://www.ncbi.nlm.nih.gov/pubmed/17393485>
25. Kim KK, Nam J, Mukouyama Y-S, Kawamoto S. Rbfox3-regulated alternative splicing of Numb promotes neuronal differentiation during development. *J Cell Biol*. 2013;200:443–58. <http://www.ncbi.nlm.nih.gov/pubmed/23420872>
26. Raj B, Irimia M, Braunschweig U, Sterne-Weiler T, O'Hanlon D, Lin ZY, et al. A global regulatory mechanism for activating an exon network required for neurogenesis. *Mol Cell*. 2014;56:90–103. <https://www.ncbi.nlm.nih.gov/pubmed/25219497>
27. Norris AD, Gao S, Norris ML, Ray D, Ramani AK, Fraser AG, et al. A pair of RNA-binding proteins controls networks of splicing events contributing to specialization of neural cell types. *Mol Cell*. 2014;54:946–59. <http://www.ncbi.nlm.nih.gov/pubmed/24910101>
28. Pimentel H, Parra M, Gee S, Ghanem D, An X, Li J, et al. A dynamic alternative splicing program regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2014;42:4031–42. <http://www.ncbi.nlm.nih.gov/pubmed/24442673>
29. Pimentel H, Parra M, Gee SL, Mohandas N, Pachter L, Conboy JG. A dynamic intron retention program enriched in RNA processing genes regulates gene expression during terminal erythropoiesis. *Nucleic Acids Res*. 2016;44:838–51. <http://www.ncbi.nlm.nih.gov/pubmed/26531823>
30. Patro R, Mount SM, Kingsford C. Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. *Nat Biotechnol*. 2014;32:462–4. <http://www.ncbi.nlm.nih.gov/pubmed/24752080>
31. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017; <http://www.ncbi.nlm.nih.gov/pubmed/28263959>
32. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol*. 2016;34:525–7. <http://www.ncbi.nlm.nih.gov/pubmed/27043002>
33. Brown JWS, Calixto CPG, Zhang R. High-quality reference transcript datasets hold the key to transcript-specific RNA-sequencing analysis in plants. *New Phytol*. 2016; <http://www.ncbi.nlm.nih.gov/pubmed/27659901>
34. Zhang R, Calixto CPG, Tzioutziou NA, James AB, Simpson CG, Guo W, et al. AtrTD—a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in *Arabidopsis thaliana*. *New Phytol*. 2015;208:96–101. <https://www.ncbi.nlm.nih.gov/pubmed/26111100>
35. Zhang R, Calixto CPG, Marquez Y, Venhuizen P, Tzioutziou NA, Guo W, et al. A high quality *Arabidopsis* transcriptome for accurate transcript-level analysis of alternative splicing. *Nucleic Acids Res*. 2017;45:5061–73. <http://www.ncbi.nlm.nih.gov/pubmed/28402429>
36. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods*. 2018; 15(3):201–6. <https://www.ncbi.nlm.nih.gov/pubmed/29334379>
37. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math North-Holland*. 1987;20:53–65.
38. Sebestyén E, Singh B, Miñana B, Pagès A, Mateo F, Pujana MA, et al. Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. *Genome Res*. 2016;26:732–44. <http://www.ncbi.nlm.nih.gov/pubmed/27197215>
39. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*. 2011;12:323. <http://www.ncbi.nlm.nih.gov/pubmed/21816040>
40. Zhang R, Lahens NF, Ballance HI, Hughes ME, Hogenesch JB. A circadian gene expression atlas in mammals: implications for biology and medicine. *Proc Natl Acad Sci U S A*. 2014;111:16219–24. <http://www.ncbi.nlm.nih.gov/pubmed/25349387>
41. Cole BS, Tapescu I, Allon SJ, Mallory MJ, Qiu J, Lake RJ, et al. Global analysis of physical and functional RNA targets of hnRNP L reveals distinct sequence and epigenetic features of repressed and enhanced exons. *RNA*. 2015;21: 2053–66. <http://www.ncbi.nlm.nih.gov/pubmed/26437669>
42. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol*. 2013;14:R36. <http://www.ncbi.nlm.nih.gov/pubmed/23618408>
43. Best A, James K, Dalglish C, Hong E, Kheirolah-Kouhestani M, Curk T, Xu Y, Danilenko M, Hussain R, Keavney B, Wipat A, Klinck R, Cowell I, Lee KC, Austin C, Venables JP, Chabot B SKM, Tyson-Capper A, et al. Investigation into human Tra2 protein-dependent splicing in MDA-MB-231 cells using iCLIP and RNA-seq. *Gene Expression Omnibus (GEO)*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59335>
44. Zhang R, Lahens NF, Ballance HI, Hughes ME HJ. A circadian gene expression atlas in mammals assayed by RNA-seq. *Gene Expression Omnibus (GEO)*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE54651>
45. Buskamp V LN. Rapid neurogenesis through transcriptional activation in human stem cell. *Gene Expression Omnibus (GEO)*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60548>
46. Pimentel H, Parra M, Gee S, Ghanem D, An X, Li J, Mohandas N, Pachter L CJ. RNA-seq analysis of differentiating human erythroblasts. *Gene Expression Omnibus (GEO)*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53635>
47. Trincado J, Entizne J, Skalic M, Eyras E. SUPPA 2.2. 2018. <https://doi.org/10.5281/zenodo.1173727>

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

